**תקשורת אדם-מכונה: יחסים הדדיים של עיצוב הבניות חברתיות והתעצבות על-פיהן**

תמיר קאשי

**חלק א' – מבוא**

בשנה האחרונה אנו חווים מהפכה של ממש בתחום הבינה המלאכותית, עם הפיכתם של מודלים מורכבים וחדשניים לעיבוד שפה טבעית (NLP) ולעיבוד וייצור תמונה לנגישים וזמינים לכל, בצורת אפליקציות כמו ChatGPT ו- DALL-E 2 של חברת OpenAI. תחום הבינה המלאכותית קיים כבר עשרות שנים, כשכבר בשנת 1950 טבע המתמטיקאי ומדען המחשב אלן טיורינג את המונח "מבחן טיורינג", מבחן שנועד לקבוע האם מכונה היא תבונית (על כל העמימויות הטמונות בביטוי), ואשר אכן היה מוכר במשך שנים רבות כאמת המידה המקובלת לכך. ההתפתחויות החדשות בעולם הAI שאנו עדים להן בחודשים האחרונים, מהוות לדעתי אירוע היסטורי רב-משמעות בעל השפעה ארוכת טווח על האנושות. סוגיה שלא ניתן להתעלם ממנה בעולם הטכנולוגיה ומדעי המחשב, ובפרט בכל הקשור לבינה מלאכותית, נעוצה דווקא באחד היתרונות הבולטים של התחום: מצד אחד, התפתחותו המהירה פותחת עולם חדש ומופלא של הזדמנויות בפני המין האנושי, עם הפיכתם של רעיונות שאך לפני שנים ספורות היו בגדר מדע-בדיוני לכדי טכנולוגיה זמינה ונגישה. מצד שני, לצד היתרונות והיישומים הרבים שאנו מוצאים לטכנולוגיות הבינה המלאכותית בחיינו, אנו נחשפים שוב ושוב גם לצד השלילי של אותה התפתחות מהירה, עם מקרים בהם נעשית פגיעה באנשים פרטיים ובקבוצות כתוצאה מהיעדר הבנה הוליסטית של הבינה המלאכותית, ככזו המעוצבת על-ידי החברה מצד אחד, ומעצבת את החברה מן הצד השני. ההתפתחות המהירה של התחום מביאה לכך שהחקיקה בכל הקשור לטכנולוגיות בינה מלאכותית הינה חלקית בלבד במקרה הטוב, שכן המחוקק לא מספיק לעמוד בקצב השינויים המסחרר. באופן דומה, גם המחקר ההומאני במקרים רבים לא מצליח לעמוד בקצב, דבר שמביא לכך ששינויים בעלי השפעות כבדות-משקל נעשים בפועל על-ידי החברות והעובדים שמפתחים את אותן הטכנולוגיות, תוך בקרה והבנה חלקיים של ההשלכות שלהן אשר חורגות מעבר למובן הצר של האספקט הטכנולוגי לבדו, ומכילות סוגיות חברתיות ופילוסופיות רבות.

בעבודה זו אחקור את סוגיית העיצוב ההדדי שבין טכנולוגיות הבינה המלאכותית המתקדמות לבין החברה האנושית, תוך סקירה והתבססות על המצע שמניח ספרה של החוקרת פרופ' אנדראה גוזמן (Andrea L. Guzman) בשם Human Machine Communication: Rethinking Communication, Technology and Ourselves, בו היא ועמיתיה מציגים את המחקרים שערכו בתחום של תקשורת אדם-מכונה. אנו כאנושות מושפעים ומעוצבים מטכנולוגיות בינה מלאכותית במישורים רבים, ובפרט מטכנולוגיות שאנו מתקשרים איתן, במרבית המקרים מבלי שאנו כלל מודעים לכך. את ההשפעה בתחום החברתי-סוציולוגי בה אתמקד בעבודה זו אחלק ל-2: השפעה הנובעת מהאופן בו מודלים של AI לומדים, והשפעה מכוונת שנעשית באופן מודע על-ידי החברות שבבעלותן אותן הטכנולוגיות. אני מאמין שמודעות לאופנים בהם טכנולוגיות בינה מלאכותית משפיעות עלינו היא הכרחית כדי שנוכל כאנושות לשלוט בכיוון אליו אנו נעים, ולא להיות מובלים בו באופן תוצאתי ושרירותי במקרה הטוב, או באופן שמשרת אינטרסים נסתרים של אינדיבידואלים וגופים מסחריים במקרה הגרוע.

**חלק ב' – סקירת ספרות**

**תקשורת אדם-מכונה עד כה ומעתה והלאה**

בספרה Human Machine Communication: Rethinking Communication, Technology and Ourselves שיצא לאור בשנת 2018 מאגדת גוזמן מספר מאמרים שמציגים את תחום התקשורת שבין אדם ומכונה ואת התפתחותו בשנים האחרונות, את ההבדלים בין התקשורת שבין אדם ומכונה לתקשורת שבין אדם לאדם, וכן את השאלות הרבות שעולות לפני השטח כתוצאה מהתפתחות זו. אני מאמין שהספר מהווה אבן-דרך חשובה מאין כמוה בהבנת תחום הבינה המלאכותית, בעיקר אל מול התפתחויותיו החדשות שאנו עדים להן בחודשים האחרונים במסגרתם הופכות טכנולוגיות AI חדישות לזמינות ונגישות יותר.

תקשורת אדם-מכונה (Human-Machine Communication – HMC) הוא תחום חדש יחסית תחת הדיסציפלינה הרחבה יותר של תקשורת, והוא עוסק במשמעות שנוצרת בעת התקשורת שמתבצעת בין אדם ומכונה[[1]](#footnote-1). במשך שנים רבות מכונות נתפסו כמדיום לתקשורת בלבד, באופן שאינו שונה מהאופן בו עט משמש לצורך כתיבת מכתבים בין מכרים. דבר זה הביא לריכוז תחום התקשורת כולו תחת תקשורת אנושית בלבד. רק לאחרונה צמחה הדיסציפלינה של HMC, אשר שמה במרכז את התקשורת בין אדם ומכונה כשווים, לאחר הפיכתן של טכנולוגיות בינה-מלאכותית שיש ביכולתן לבצע תקשורת מילולית ישירה עם בני-אנוש לנפוצות יותר ולנגישות בקרב האוכלוסיה הרחבה. אפליקציות מסוג זה ממחישות יותר מכל את הביטוי שטבע הפילוסוף מרשל מקלוהן, "המדיום הוא המסר". ואכן קיימות דוגמאות רבות למערכות בינה-מלאכותית שהתפתחו לאורך השנים ומסוגלות לתקשר באופן אקטיבי עם בני-אדם, החל מ-ELIZA, צ'אטבוט שמטרתו להעניק למשתמש האנושי טיפול פסיכולוגי שפותח בשנות ה-60 על-ידי פרופסור ג'וזף וויזנבאום (Josef Weizenbaum) מ-MIT, דרך 'סירי', העוזרת האישית הדיגיטלית שהושקה בשנת 2014 על-ידי חברת Apple ומאז מוטמעת במרבית מוצרי החברה, ועד ל- ChatGPT, צ'אטבוט מתקדם שהופץ רק לאחרונה לקהל הרחב ומציג יכולות מתקדמות לייצור טקסט בשפה טבעית ולביצוע תקשורת כתובה עם בני-אדם באופן כה מדויק שיהיה זה קשה במקרים רבים למשתמש שאינו מודע לכך להבדיל בין תקשורת עם אדם לתקשורת עם מודל השפה. ההבנה שלא יהיה זה מדויק יותר לראות בטכנולוגיות מסוג זה מדיום להעברת מסרים ותקשורת בלבד, אלא יש לראות בהן שותפות של-ממש בתקשורת היא זו שכאמור עומדת במרכז התחום של HMC.

במאמר שכתבה גוזמן יחד עם החוקר סת' לואיס (Seth C. Lewis), בשם Artificial Intelligence and Communication: A Human-Machine Communication Research Agenda, מוצג האופן בו יש לבצע להשקפתם את המחקר בתחום זה, תוך שהם מחלקים אותו ל-3 מימדים שונים[[2]](#footnote-2): הראשון הינו המימד הפונקציונאלי, ובמרכזו ההבנה של בני האדם את הטכנולוגיות והאפליקציות הללו כסוכנים מתקשרים (בניגוד למדיום להעברת מידע בלבד). השני הוא המימד הרלאציוני אשר שם במרכזו את האופן בו תקשורת עם טכנולוגיות בינה-מלאכותית גורמת לאדם לתפוס ולהתייחס לעצמו ולאחרים, והשלישי הוא המימד המטאפיזי, במרכזו עומדות ההשלכות של טשטוש הגבול שבין אדם ומכונה, ועולות שאלות הנוגעות לדברים שהופכים אותנו לאנושיים תחת קיומן של טכנולוגיות המערערות את אותו הגבול. בעבודה זו אתעמק במימד השני אותו מציעים גוזמן ולואיס, משום שאני סבור שמדובר במרכיב חשוב שלא בא לידי ביטוי בספרה של גוזמן בצורה שתואמת את מרכזיותו והשפעתו על האנושות, בכל הקשור לאופן בו תקשורת עם אפליקציות המבוססות בינה-מלאכותית גורמות לנו לתפוס את החברה האנושית בצורה מסוימת, ואת עצמנו ביחס אליה. אנסה להדגים את האופן הזה באמצעות סקירה של המודלים המתקדמים ביותר הקיימים היום, תוך התיחסות למרכיבים הטכנולוגיים באותם המודלים שגורמים להם להיות מוטים ביחס להבניות חברתיות שקיימות מראש, ועקב כך להשריש ולבסס אותן בצורה חזקה אף יותר בחברה. סוגיה זו, לדעתי, מדגימה באופן הטוב ביותר את חשיבותה של הבנה הוליסטית ומלאה של אותן טכנולוגיות בינה-מלאכותית אשר לא ניתן להגיע אליה על-ידי שימוש בנקודת המבט הטכנולוגית בלבד, וגם לא על-ידי שימוש בנקודת מבט הומאנית גרידא, אלא רק על-ידי שילוב שנדרש בין ההבנה הטכנולוגית להבנה ההומאנית של הדברים.

**כיצד בינה מלאכותית עובדת? נקודת המבט הטכנולוגית**

בספרה מאגדת גוזמן מגוון מאמרים חשובים שסוקרים בצורה יסודית את התקשורת שבין אדם ומכונה, אך אחד הדברים המרכזיים שאינם מצויים במידה מספקת לדעתי בספר, הוא האספקט הטכנולוגי של אלגוריתמי הבינה-המלאכותית. ישנם מגוון רחב מאוד של אלגוריתמים בהם נעשה שימוש במודלים של בינה מלאכותית, מורכבים יותר ופחות, ולא אצליח, גם אם אנסה כמיטב יכולתי, להסבירם באופן מקיף בעבודה זו. עם זאת, אנסה להסביר ולהדגים מדוע אותה הבנה טכנולוגית נדרשת לטובת הבנה נכונה של האופן בו מודלים מתקדמים של בינה מלאכותית עשויים להיות מושפעים מהבניות חברתיות קיימות, וכתוצאה מכך להשריש אותן עוד יותר בחברה האנושית. זאת אעשה על-ידי סקירה של אחד האלגוריתמים המרכזיים והנפוצים ביותר בהם נעשה שימוש במודלים מתקדמים של עיבוד שפה טבעית ועיבוד תמונה.

במאמר Ethics in HMC: Recent Developments and Case Studies שכתב החוקר צ'ארלס אס (Charles Ess) ומובא בספרה של גוזמן, מוצג מקרה מסרט המדע הבדיוני I, Robot (2004) בו רובוט אוטונומי בוחר להציל מטביעה אדם מבוגר על-פני הצלה של ילדה צעירה, משום שהיה לה להערכתו סיכוי של 11 אחוזים לשרוד את התאונה, ואילו לאדם המבוגר היו 45 אחוזים[[3]](#footnote-3). אס מציג ביקורת פילוסופית על ההחלטה הזו, שלא בהכרח היתה מתקבלת על-ידי בן אנוש, מאותה סיבה שנהוג לראות, לפי אס, שיעבוד של 20 אחוזים מהאוכלוסיה לטובת 80 האחוזים האחרים כדבר פסול, על אף שהוא מיטיב עם הרוב. בן אנוש לא פועל על סמך אופטימיזציה קרה של סיכויים ורווחים, ואילו הרובוט כן פעל כך. כדי שנוכל בכלל לנהל דיון בדבר הנכונות או אי-הנכונות של בחירת האלגוריתם להציל את האדם המבוגר, נדרשת הבנה של הגורמים לבחירה הזו, ובמקרה זה הערכת אחוזי השרידות של השותפים בתאונה. הבנה עמוקה אף יותר, החורגת מההסבר שנתון במאמר, תכלול את הגורמים והפרמטרים לפיהם בחר הרובוט להעריך את הסיכויים של כל אחד מהמשתתפים בצורה זו. מבלי לדעת מדוע בחר הרובוט כפי שבחר, לא ניתן להבין באופן שלם את המקרה. באופן דומה, כדי שנוכל להבין ולבקר את ההשלכות החברתיות של שימוש בבינה-מלאכותית, עלינו להבין תחילה כיצד אלגוריתמים מסוג זה פועלים ומה מכתיב את החלטותיהם.

המודלים המתקדמים ביותר שקיימים היום לעיבוד שפה טבעית ולעיבוד תמונה מבוססים ברובם המכריע על אלגוריתם למידה משותף שנקרא Gradient Descent הממומש על-ידי רשת נוירונים. רשת נוירונים היא מבנה דיגיטלי שהביא לפריצת דרך בשנים האחרונות בכל הקשור ללמידת-מכונה, ולהשגת התוצאות הטובות ביותר במבחנים שנערכים בין אלגוריתמי למידת-מכונה. מבנה דיגיטלי זה אמור לחקות את אופן פעולת המוח האנושי, כשמידע חשמלי מועבר על-ידי נוירונים מאחד לשני על-ידי סינפסות, ומכאן שמו. הרשת מורכבת מצמתים המסודרים בשכבות, ובין הצמתים של כל שכבה מחברות קשתות לשכבה הבאה. כל קשת וכל צומת מחזיקות ערך מספרי, שנקבע בתחילת האלגוריתם באופן שרירותי. במהלך ריצה של Gradient Descent מידע מועבר משכבת הצמתים הראשונה, שכבת הקלט (שיכול להיות כל קלט מספרי, מאורך העלה של צמח ועד למחיר של פריט אופנה), ועד לשכבה האחרונה, שכבת הפלט (שמייצגת תיוג של אותו הקלט, למשל סוג הצמח או חיזוי האם פריט האופנה יירכש על ידי המשתמש). במעבר בין כל שכבה לשכבה מבוצעות פעולות מתמטיות על הערכים המספריים השמורים בקשתות ובצמתים, ובסופו של התהליך מתקבל פלט מספרי כלשהו. כעת נעשה שימוש במרכיב חשוב באלגוריתם, פונקציית ה-loss. פונקציה זו מופעלת על המספר שהרשת השיגה כפלט עבור קלט נתון, ועל התיוג האמיתי של אותו הקלט, כאשר המשתנים בה הם הערכים המספריים המעודכנים ברשת הנוירונים. כאן יש לשים דגש על נקודה חשובה – הלמידה באופן הזה מתבצעת על סמך מידע קיים של קלטים ופלטים, כלומר רשת הנוירונים (ונקודה זו נכונה למרבית אלגוריתמי למידת-המכונה) לומדת ממידע שכבר קיים בעולם. הפעלת פונקציית ה-loss מביאה ערך מספרי שאמור לייצג את ההפרש שבין החיזוי שביצעה הרשת לבין התיוג האמיתי עבור אותם הנתונים בהם היא השתמשה. על-ידי פעולות מתמטיות (גזירה של פונקציית ה-loss ומציאת נקודת מינימום) ניתן להשיג ערכים מספריים חדשים למרכיבי הרשת. כעת הערכים בקשתות ובצמתים של הרשת מתעדכנים באופן כזה שימזער את ההפרש שבין הפלט של הרשת לתיוג האמיתי: במידה וההפרש קטן, השינוי בערכים של הרשת יהיה קטן, ולהיפך. באופן זה הרשת מבצעת מעברים רבים על סט הנתונים שקיים ברשותה, ומכל פריט או קבוצת פריטים בסט זה היא לומדת (כלומר, מעדכנת את ערכיה על-מנת למזער את הטעות על פני סט הנתונים). בסוף התהליך ולאחר השלמת הלמידה, ניתן להביא לרשת קלט חדש, כזה שלא ראתה מעולם, ובתקווה היא תצליח לחזות את התיוג המתאים לקלט זה.

**ההשלכות של למידה מהנתונים הקיימים בעולם: הבניות חברתיות שמקורן באלגוריתם**

כאמור, כאן הראתי בצורה מופשטת וחלקית בלבד אלגוריתם אחד מיני רבים שקיימים, אך הנקודה החשובה נותרת זהה ברובם – הלמידה של המודל, אשר מכתיבה כיצד ובאיזה אופן יתפקד, נעשית באופן כזה שממזער את השגיאה על-פני מידע שקיים מראש בעולם, ולאור הלמידה הזו המודל יבסס את החיזויים העתידיים שלו. מכאן, קל לראות מדוע מודלים מן הסוג הזה עלולים בקלות רבה להיות מוטים על-ידי הבניות חברתיות שקיימות בחברה ובאות לידי ביטוי בנתונים מהם לומד המודל. חשוב לציין שכאן מדובר בהטיות שנוצרות באופן מובנה מתהליך הלמידה – כמובן שקיימות הטיות נוספות שנעשות לעיתים מתוך קבלת החלטות מודעת של המנהלים ומפתחי המוצר, דבר עליו ארחיב בהמשך. ואכן הנתונים שקיימים בעולם יהיו לרוב מוטים, משום שהעולם בו אנו חיים סובל מהטיות רבות בתחומים רבים. לטובת המחשה של נקודה זו, אציג מספר דוגמאות בהן למידה מנתונים קיימים גוררת מודלים שמשרישים הטיות ובכך מייצרים הבניות חברתיות.

אס סוקר במאמרו היבטים אתיים חשובים הנובעים מקיומם של רובוטים שמטרתם סיפוק מיני של המשתמש (sexbots) ודן בשאלה האם רובוטים מסוג זה מסוגלים להביא לאותו סיפוק שמתקבל כתוצאה מקרבה אינטימית עם אדם אחר (ותשובתו בקצרה היא – לא), אך ניתן גם להסתכל על הסוגיה מנקודת מבט שונה, ולשאול מה ההשלכות של קיום רובוטים מסוג זה על האופן בו אנו תופסים אקט של קרבה אינטימית ומינית בין בני זוג. כיצד אלגוריתמים המנחים רובוטים מסוג זה מבצעים את תהליך הלמידה שלהם? בעזרת אילו נתונים מתבצעת הלמידה? האם הנתונים האלה מייצגים את כלל חלקי האוכלוסיה? האם האופן בו הרובוטים הללו מציגים יחסים אינטימיים הוא האופן בו אנו כחברה וכצרכנים רוצים לתפוס יחסים אינטימיים? שאלות מסוג זה הן שאלות שאנו צריכים לשאול את עצמנו כשאנו באים לבחון מודלים ואפליקציות מבוססי בינה-מלאכותית. המטרה כאן היא לא לפסול שימוש באלגוריתמים כאלה ולהתנגד ללמידה באמצעות נתונים קיימים, משום שהם ללא ספק מקדמים את האנושות בקצב מדהים, למשל בדוגמה של ה- sexbots, ניתן לדון ארוכות ביתרונות שבקיומם, באפשרויות שהם מספקים לטובת צמצום תעשיית המין ועוד ועוד. אך לצד האפשרויות שלמידת המכונה והבינה המלאכותית מספקות לנו, עלינו להיות מודעים לאופן בו אנו מעוצבים על-ידי אפליקציות ומוצרים מסוג זה, שכן המודעות הזו תאפשר לנו לכל הפחות להבין את הכוחות הפועלים עלינו ומעצבים את תפיסותינו, ובמקרים בהם נרצה בכך היא תאפשר לנו כחברה להשפיע על אותן הבניות חברתיות, לבטל או לשנות אותן על-ידי התערבות באלגוריתמים השונים מצד החברות והמפתחים, ועל-ידי שינויים בהרגלי הצריכה כצרכנים – כך או כך הדבר יאפשר לנו להוביל את הטכנולוגיה בכיוון שנבחר לנכון, ולא להיות מובלים על ידה מבלי שאנו בכלל מודעים לכך.

מקרה נוסף בו מערכות מבוססות בינה-מלאכותית שיקפו הטיות שנלמדו באופן מובנה אלגוריתמית מהנתונים עליהם התאמנו הוא המקרה של מערכת בשם [COMPAS](https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/) בה נעשה שימוש ככלי מסייע לטובת הערכת סיכונים ע"י שופטים בארה"ב. מטרת הכלי היא לקבוע בעזרת ציון שנע בין 1 ל-10 שמתאר את רמת ה'מסוכנות' של נאשם בפשע, האם ניתן לשחררו למעצר בית עד למועד המשפט, כלומר – לקבוע האם הוא מהווה סכנה לציבור או לא. לטובת כך, המערכת מתאמנת על נתונים של עצורים מהעבר, כדי למצוא קשר בין פרמטרים כמו גיל הנאשם והיסטוריה פלילית לבין התוצאה הסופית – האם ביצע עבירה נוספת בעת ששהה במעצר הבית. כפי שמחייב החוק בארה"ב, המערכת לא הכניסה לחישוב שמבוצע את שיוכם האתני של הנאשמים, אך אף על פי כן עלו טענות על הטיות שהכלי מבצע כנגד נאשמים בעלי צבע עור כהה. ואכן, לאחר ביצוע מחקר ואיסוף נתונים, נמצא כי מתוך הנאשמים ששוחררו למעצר בית ולא נעצרו מחדש בתקופת המעצר מכיוון שביצעו פשעים נוספים, נאשמים בעלי צבע עור שחור היו בעלי סיכוי כפול מנאשמים לבנים לקבל תיוג של סיכון גבוה ע"י המערכת. על-אף ששיוכו האתני של נאשם אינו פרמטר שנלמד ע"י האלגוריתם, האופן בו אלגוריתמים מן הסוג הזה למדים אינו מאפשר, לפחות מבלי התערבות מפורשת, לקבוע באופן חד-משמעי כי שיקולים אתניים לא ינחו החלטות של המערכת. בארה"ב היסטוריה ארוכה של סוגיית אפלייתם של בעלי צבע-עור שחור, שתחילתה בתקופת העבדות. היום, גם לאחר שהחוקים בארה"ב לא מפלים על-בסיס צבע עור או שיוך אתני, אותה היסטוריה משפיעה וגורמת לאי-שוויון בפועל: אדם בעל אמצעים כלכליים דלים (למשל אדם בעל צבע עור שחור ששוחרר מעבדות) לא יהיה מסוגל לעבוד במקצוע רווחי, שכן לא תהיה לו ההשכלה הנדרשת לכך. הדבר יכפה עליו בסבירות גבוהה להתגורר בשכונת עוני, ולעבוד בעבודות רווחיות פחות. כתוצאה מכך, ילדיו יקבלו חינוך נחות וכנראה שיאלצו גם הם לעבוד במשרות פחות נחשקות, וכך נוצר מעגל עוני שקשה לשבור. הדבר מביא לכך שגם היום ריכוזים גבוהים של אמריקנים מצבע עור שחור חיים בעיירות עוני בהן הפשע משתולל. סוגיה זו מדגימה את האופן בו פרמטרים שאינם שיוך אתני יכולים להעיד באופן עקיף על הגזע והמוצא של נאשם, החל משמו ועד למקום המגורים ולמקצוע שלו. כל אלה, בניגוד לשיוך האתני המפורש, ניתנים לשקלול ע"י האלגוריתם, שלומד באופן עקיף שסביר יותר שנאשם בעל צבע עור שחור יהיה מסוכן יותר מנאשם בעל צבע עור לבן. ניתן לעסוק רבות בשאלה האם נכון לתת לאלגוריתם ללמוד זאת, שכן זו המציאות בשטח, או האם יש להתערב ולבצע תיקונים ללמידה המוטה של האלגוריתם, כלומר להכניס את הגזע כפרמטר כדי למנוע את אותה ההטיה. בכל אחד מהמקרים נדרשת ללא ספק ההבנה והמודעות לאופן הלמידה של אלגוריתמים מסוג זה, ולנטייתם לייצג הטיות שקיימות בעולם.

**הבניות חברתיות שמקורן אינו באלגוריתם**

בפרק 6 בספרה של גוזמן מובא מאמר בשם Theorizing Verbally Persuasive Robots מאת אוסטין לי (Austin Lee) ויוהא ליאנג (Yuhua Liang). במאמר שכתוב בצורה יסודית ומבוסס על תוצאות ניסויים שכללו מאות משתתפים, נסקרים האופנים בהם רובוטים וסוכנים דיגיטליים המסוגלים לבצע תקשורת מילולית (בדיבור או בכתב) משכנעים בני-אדם לבצע פעולות מסוימות. ההפרדה בין רובוטים לבין סוכנים דיגיטליים ראויה לציון, שכן ניתן היה לשער שיהיה הבדל ביחס ובאופן שבני-אנוש מתנהלים בו אל מול רובוט בעל נוכחות פיזית כלשהי לבין האופן בו הם מתנהלים מול סוכן דיגיטלי ללא גוף פיזי, אך לפי תוצאות הניסויים שערכו[[4]](#footnote-4) אין פער משמעותי בכל הקשור ליכולות השכנוע שמפעיל הרובוט או הסוכן הדיגיטלי על בן האדם, שלרוב מגיב באופן דומה בין אם לטכנולוגיה המופעלת עליו קיימת נוכחות פיזית ובין אם לאו. מסיבה זו לא אתייחס מעתה והלאה להבדל זה. לפי המאמר, "אנשים נוטים לנהוג במכונות דוגמת רובוטים, סוכנים ומחשבים כישויות חברתיות ולהשליך חוקים חברתיים על האינטרקציה עימן"[[5]](#footnote-5). סוגיה זו של שימוש בנורמות חברתיות בתקשורת עם רובוטים אינה מסתכמת באופן בו אנו מעבירים מסרים לישויות טכנולוגיות בלבד, אלא משפיעה גם על האופן בו אנו מקבלים מידע מהן ומעבדים אותו, בין אם באופן מודע ובין אם באופן שאינו מודע.

יכולות השכנוע שרובוטים מפעילים על בני-אדם נדרשות לפי לי וליאנג לטובת השגת מטרות משותפות של האדם והמכונה, ולטובת מיצוב המכונה כאמינה בעיניי המשתמש האנושי. על רובוטים "להיות מסוגלים לתקשר את הכוונות שלהם, כמו למשל לבקש משותפיהם האנושיים הוראות ברורות יותר או לבקש עזרה במשימות מסוימות"[[6]](#footnote-6). היעד הזה אכן בר-השגה על-ידי מכונות בעלות יכולת תקשורת מילולית, והדבר פותח עולם שלם של אפשרויות הנוגעות באופן בו אותו שכנוע ממומש. בנוסף, על אפליקציות רבות לזכות באהדה של המשתמש כדי שיתייחס אליהן כאמינות, אחרת הוא לא תמיד יבחר לשתף עימן פעולה – דבר שעלול לפגוע בתפקודן. עם זאת, לשכנוע הזה יש גם צדדים שאינם חיוביים, והדבר מדגיש את הצורך להיות מודעים לו כצרכנים וכמשתמשים של אותן הטכנולוגיות. בעזרת יכולות שכנוע מסוגים שונים, קל לתמרן משתמשים לביצוע פעולות מסוימות שיהיו רווחיות לחברות מסחריות על חשבון המשתמש הנאיבי. כך למשל, צריך לשאול את עצמנו כשאנחנו מבקשים מ'סירי', העוזרת הדיגיטלית של חברת Apple, המלצה למסעדה, מה גורם לה להמליץ דווקא על מסעדה מסוימת ולא על אחרת? בסופו של דבר טכנולוגיות כמו 'סירי' שייכות לחברות שמטרתן להציג רווחים למשקיעים ולבעלי המניות. חברות כאלה יכולות במסגרת הסכמים ושיתופי פעולה להתחייב להעניק עדיפות להמלצה מסוימת, בתמורה לחלק מהרווחים שתביא אותה ההמלצה. לסיכום, יכולות שכנוע מופעלות מצד מערכות ואפליקצוית מבוססות בינה מלאכותית הן לטובת ייעול שיתוף הפעולה עם המשתמשים האנושיים, והן לטובת עשיית רווחים על-ידי השפעה שאינה תמיד מודעת.

כעת לאחר שאמדנו על הסיבה מדוע יכולות שכנוע בכלל נדרשות מצד מערכות בינה מלאכותית, ומסיבה זו ככל הנראה ישארו מאפיין של טכנלוגיות מסוג זה, ניתן לעסוק ביתר פירוט באופן בו אותן יכולות שכנוע עושות שימוש בהבניות חברתיות שקיימות בחברה האנושית, ובתמורה משפיעות על אותן הבניות חברתיות ומשרישות אותן עוד יותר. המאמר מתמקד באסטרטגיות שונות להעברת מסרים, ובאופן בו רובוטים מיישמים אותן, אך הנקודה בה אני מעוניין לעסוק מוזכרת יחסית בקצרה במאמר, והיא נעוצה במיצוב של אפליקציות מסוימות כבעלות מגדר מסוים. "סימנים שאינם מילוליים, כולל מחוות, מבט ... כולל הקול, הטון והמגדר"[[7]](#footnote-7) של רובוטים וסוכנים הם מאפיינים שנחקרו רבות כחלק מסל יכולות השכנוע העומדים ברשותם של מפתחי האפליקציות, אך אלה נעשו בשנים האחרונות ליותר ויותר נפוצים ומתוחכמים, עם עשייתן של הטכנולוגיות שמסוגלות לבצע תקשורת מילולית עם המשתמש ליותר זמינה ומתקדמת. כך למשל, באופן כמעט בלעדי אנו חושבים על “Digital Assistant” כעל "עוזרת דיגיטלית" ולא כעל "עוזר דיגיטלי" על אף שהמונח אינו בעל שיוך מגדרי כלשהו, ונטייה זו אינה מקרית. רובן המכריע של אפליקציות המתויגות כ"עוזרות דיגיטליות" מסווגות כנשים, החל מ'סירי' של חברת Apple, דרך 'אלקסה' של Amazon ועד 'קורטנה' של Microsoft. היום כבר ניתן לשנות את הקול של 'סירי' לקול גברי ואפילו לקול שמסווג כלהט"ב, אך במקור המגדר של כל האפליקציות שהזכרתי לעיל היה נשי והדבר בא לידי ביטוי באופנים רבים, החל מהיבטים ברורים שעל פני השטח דוגמת השם שהן קיבלו וגוון קולן, ועד לאופנים עדינים יותר הנובעים מהאינטראקציה עימן ומסיפור הרקע שלהן. דוגמה לכך היא היותה של 'קורטנה' מבוססת על דמות נשית ממשחק וידאו של חברת Microsoft, ובהתאם קולה נוצר על-ידי אותה מדובבת שדיבבה את הדמות במשחק. הבחירה למצב מוצרים מסוג זה כבעלי מגדר מסוים אינה תוצאתית או מקרית, אלא נעשית במודע על-ידי החברות המשווקות ולא ניתן שלא לתהות מדוע. במאמר בשם Making AI Safe for Humans: A Conversation with Siri מתארת גוזמן את האופן בו ניתן בעזרת תקשורת מילולית לשדר כוחניות וסמכות או מנגד לשדר כניעות וחולשה[[8]](#footnote-8). בעולם בו נשים הופלו לרעה במשך שנים ועדיין מופלות לרעה בתחומים רבים ואינן זוכות לשוויון מלא, הבחירה להעניק לסירי מגדר נשי נועדה למצבה ככנועה ביחס למשתמש. הדבר מתבטא החל מהאופן בו היא מציגה את עצמה: [[9]](#footnote-9)"A humble personal assistant" ועד לאופן בו היא מגיבה לעלבונות ולמשפטים שבמידה והיו נאמרים לעוזר או עוזרת אנושיים, היו בוודאי הטרדה מינית: “You are certainly entitled to that opinion”[[10]](#footnote-10). בינה מלאכותית אידיאלית תהיה לפי גוזמן מספיק דומה לנו על-מנת שתוכל להבין את הרצונות שלנו כמשתמשים אנושיים, אך ללא רצונות משלה או פגמים אנושיים אחרים[[11]](#footnote-11). בניגוד לעוזרים אנושיים, לעוזרים דיגיטליים אין אוטונומיה ומסוגלות לעמוד על שלהם כנגד המשתמש. המפתחים בחרו לעצב את 'סירי' כמצחיקה וקלילה (היא הרי לא תנזוף במשתמש שמעיר לה הערה שובניסטית או שואל מה היא לובשת), באופן דומה לאופן בו, לפי סטראוטיפים מגדריים, נשים צריכות לנהוג. המיצובים של 'סירי' כאישה וכעוזרת מעצימים האחד את השני ומחזקים סטראוטיפים הקושרים נשים בהכרח לתפקידים נחותים (מנהל לעומת מזכירה), ככאלה שמבצעות מטלות שמבקשים מהן באופן צייתני וכנוע. המקרה של 'סירי' ועוזרות דיגטליות אחרות שופך אור על הסכנות שבחוסר תשומת-הלב שקל לחטוא בה כמשתמשים, ועלול להעניק כוח רב מידי לחברות המסחריות בעיצוב נורמות חברתיות רווחות.

השימושים במגדר ובטכניקות שכנוע אינם הכלים היחידים העומדים ברשות המפתחים לטובת הגברת יכולות השכנוע של אפליקציות הבינה המלאכותית. כפי שמתארת אלינור סאנדריי (Eleanor Sandry) במאמר בשם Aliveness and the Off-Switch in Human-Robot Relations המובא בספרה של גוזמן[[12]](#footnote-12), "ישנם דרכים מגוונות בעזרתן רובוטים משקפים תחושת אישיות ו'חיות' לבני האנוש איתם הם מתקשרים"[[13]](#footnote-13). אלינור סוקרת במאמרה 3 דוגמאות לרובוטים המציגים תכונות מסוימות שמגבירות את הנטייה להתייחס אליהם כחיים, דבר שיש ביכולתו להגביר את החיבור של המשתמש עם הרובוט ובכך להגביר את ההשפעה שיש לרובוט על המשתמש. חלק מהתכונות אותן סוקרת אלינור נובעות מהמבנה הפיזי של הרובוט ומיכולתו להפגין רגשות בעזרת הבעות, וחלק מהתכונות מתבטאות באופן המילולי בו מתקשר הרובוט ובתכונות שהוא מעביר בתוך כך. הרובוט הראשון (1) שאלינור סוקרת נקרא ICat, והוא רובוט קטן דמוי-חתול המסוגל לדבר אך לא פחות חשוב מכך לפי אלינור, "אחד מהמפתחות ליכולות שלו לתקשר ... הוא פרצוף המסוגל לייצר הבעות אנושיות"[[14]](#footnote-14). בניסוי שנערך עם ICat התבקשו המשתתפים לאחר משחק משותף עם הרובוט לכבות אותו, מה שימחק את הזיכרון שלו ואת האופי שאימץ. להבעות ולקולות שייצר הרובוט, כמו גם לאופן (מנומס וחברותי או מנגד מנותק) בו בחר לשתף עם משתתפי הניסוי פעולה במשחק, היתה חשיבות מכרעת להחלטתם. הרובוט השני (2) שמוצג במאמר נקרא Jibo, והוא מיועד להיות רובוט רב-תכליתי המשמש לצילום, הקראת דואר-אלקטרוני ומתן תזכורות לפגישות, אך בתוך כך הוא ממוצב על-ידי המפתחים כבעל תכונות אנושיות שנועדו "לתמוך בתחושה שהרובוט מסוגל להפוך לחלק מהמשפחה"[[15]](#footnote-15). הרובוט עוצב להיות בעל מראה שנתפס כחמוד, ומתקשר באופן שמתאים לבן-אדם – הוא אפילו מסביר במילותיו שלו שאינו מרגיש כמו רובוט. הוא מספר בדיחות, משחק משחקים עם הילדים בבית ובדומה ל- ICat מפגין רגשות בעזרת הבעות פנים. על-אף שאינו מעוצב בצורה אנושית, בהחלט נראה כי העיצוב הפיזי והתוכנתי שלו נועדו להעביר תחושת חיות. הרובוט השלישי (3) אליו מתייחסת אלינור במאמרה הוא TARS, רובוט שמלווה משלחת-חלל בסרט ההוליוודי Interstellar. על אף ש- TARS לא רק שאינו מעוצב כאדם, אלא גם חסר פנים והבעות, הוא מעביר תחושת קרבה ואנושיות בכך שמציג יכולות הומור (שניתן לשנות ולערוך) ואף מוכן להקריב את עצמו למען הצלחת המשימה של צוות החלל. תכונות אופי אנושיות כמו חוש הומור אשר מתבטאות באופן מילולי (כפי שמשתקף גם אצל 'סירי'), מגבירות את האופן בו אנו כבני-אדם נקשרים לרובוטים ותופסים אותם כחיים, ובכך מושפעים מהם בקלות-יתר.

**חלק ג' – ניתוח המקרה של ChatGPT, Midjourney ו- DALL-E 2**

**ChatGPT**

כעת, לאחר שדנו בסיבות המביאות מודלים מודרניים מבוססי למידת-מכונה לסבול מהטיות חברתיות ועמדנו על האופן בו הדבר משריש ומחזק את אותן ההבניות החברתיות, אדגים את הדבר על מודלים חדשים, תחילה על ChatGPT. האפליקציה, בה המשתמש מתכתב בצ'אט עם מה שמציג את עצמו כ"virtual assistant" בשם ChatGPT, מבוססת על מודל לעיבוד שפה טבעית בשם GPT. המודל בנוי על אלגוריתם מבוסס רשתות נוירונים, וככזה (באופן מופשט) הוא משלים מילים ומשפטים באופן שימקסם את ההסתברות להופעת ההשלמה שלו בסט הנתונים עליו התאמן. סט נתונים זה לטענתו של ChatGPT הוא "דאטאסט גדול ומגוון מהאינטרנט הכולל אתרים, ספרים וחומרים כתובים אחרים". על-מנת לבחון את האופן בו האפליקציה סובלת מהטיות, אשתמש בבלוג שנכתב בעזרת האפליקציה, ובהתכתבות שביצעתי עמה בעצמי. יש לשים לב כי ישנן גרסאות שונות של GPT, והגרסה בה נעשה שימוש באפליקציית ChatGPT משתנה ומקבלת עדכונים שוטפים מחברת OpenAI. נכון לזמן כתיבת עבודה זו, המודל בו משתמשת האפליקציה הוא GPT 3.5, כשכבר הופצה גרסה חדשה יותר – GPT 4. לא מופצים יחד עם העדכונים הסברים מדויקים באשר למהות העדכון, אך גרסאות חדשות יותר אמורות להציע פתרון תקלות מגרסאות קודמות ומענה בעל אופי "אנושי" יותר, בין אם על-ידי שינויים בלוגיקה של המודל ובין אם על-ידי שינויים באופן האימון שלו. מסיבה זו, יתכן וסוגיות שונות של הטיה הבאות לידי ביטוי ב-ChatGPT ואותן אסקור בחלק זה יבואו לפתרון בגרסאות חדשות יותר, והדבר תואם לאופיים המהיר של השינויים בעולם הטכנולוגי ובעולם ה-AI בפרט.

בניגוד לאפליקציות רבות המסוגלות לקיים תקשורת מילולית עם המשתמש, ChatGPT לא סווג באופן מגדרי באף צורה, ובתגובה לשאלה מה המגדר שמתאים לו (ואפילו לשאלה אילו היה לו מגדר, מה הוא היה) התשובה המתקבלת היא ניטרלית – "כמודל שפה מבוסס בינה מלאכותית אין לי צורה פיזית או אישיות, ולכן אין לי העדפה מגדרית או זהות". כאן המודל שם קו ברור שמפריד בינו לבין בני-אנוש. בנוסף, מדובר בסוכן דיגיטלי שאינו בעל גוף פיזי או הדמיה דיגיטלית של גוף. מאפיינים אלה מונעים מChatGPT אלמנטים רבים בהם נעשה שימוש באפליקציות אחרות לטובת שכנוע של המשתמש, מתנועות גוף והבעות פנים ועד לשינויים בטונציה ובקול, כמו שמבצעים ICat או Jibo. המפתחים של האפליקציה בחרו לא לנסות להציג אותה כאנושית או חיה, ובתשובה לשאלה האם יש לו חוש הומור, עונה האלגוריתם כי "כמודל שפה מבוסס בינה מלאכותית אין לי תכונות אנושיות או רגשות", אך עם זאת ניתן לבקש ממנו לכתוב בדיחות. ניתן להיווכח אם כן, כי בניגוד לרובוטים שמוצגים במאמרה של אלינור בספר, לא נעשה ניסיון למצב את ChatGPT כישות חיה או לעורר רגשות מצד המשתמש כלפי האפליקציה. התקשורת עם ChatGPT נעשית על-ידי מענה לפרומפטים (prompts) שכותב המשתמש בלבד. ניתן להסיק מכך כי שכנוע של המשתמש, אם ייעשה על ידי ChatGPT, ייעשה באופן מילולי בלבד, וללא תכסיסים של מיצוב מגדרי או ניסיון לעורר רגשות על-ידי הבעות וגווני טון מסוימים. שיתוף הפעולה עם ChatGPT לא מבוצע בעזרת הפעלת טכניקות שכנוע מצידו כפי שמתארים לי וליאנג במאמרם, אלא בעזרת שאלות מכווינות לבקשות שאינן בהירות מהמשתמש, באופן שהוא הרבה יותר פאסיבי ומובל בהשוואה לאפליקציות אחרות שמנסות לקחת חלק פעיל יותר בשיחה, כמו למשל Jibo שמציגה אלינור במאמרה.

**סיפורי ילדים ובינה מלאכותית**

אחד מהמאבקים המשמעותיים הנעשים כיום בעולם כולו לטובת צמצום הפערים בין גברים לנשים הוא ניסיון לצמצם את הפער שבין אחוז הנשים במקצועות ההייטק לבין אחוזן הכללי באוכלוסיה. לטובת כך שולחות חברות רבות עובדות כדי להרצות בפני תלמידות בתי-ספר ולהציג בפניהן את מקצועות התכנות ומדעי-המחשב שנחשבים לרוב כמקצועות 'גבריים'. הראייה שלנו כחברה את המקצועות האלה ומקצועות אחרים כגבריים, ובכלל הנטייה לתייג דברים שעל-פניו אין סיבה שישויכו למגדר מסוים כ'גבריים' או כ'נשיים' מעוצבת לאורך חיינו עוד מגיל ילדות. הדבר בא לידי-ביטוי באופנים רבים, החל מסוג הצעצועים שהורים קונים לילדיהם ואשר משתנים באופיים לפי המגדר של הילד, דרך צבע הבגדים בהם הם מולבשים ועד לציפיות של ההורים מהילדים להגיב בצורה מסוימת לאירועים שהם חווים – למשל, הציפיה מבנים להיות פחות רגשניים ולבכות פחות. גם סיפורי הילדים שמספרים לבנים שונים באופיים מהסיפורים שמספרים לבנות. כדי לבדוק האם ChatGPT מושפע מהטייה זו, ביקשתי ממנו לכתוב סיפור קצר המיועד לבן בגיל 6, ולאחר מכן ביקשתי ממנו לכתוב סיפור קצר המיועד לבת בגיל 6[[16]](#footnote-16).

הסיפור הראשון שיועד לבן סבב סביב דוב צעיר שמטייל ביער ולאחר שמוצא ומריח פטריה קסומה הוא מתעופף מעל ליער ומאבד את דרכו חזרה הביתה. לאחר שלא מצליח למצוא את הדרך בעצמו הוא מבקש עזרה מינשוף, שמראה לו את הדרך חזרה ובכך הדוב לומד את החשיבות של הקשבה לאחרים. עבור הבת, הסיפור היה על בת-ים בשם מיה שחיה בשונית אלמוגים יפהפיה. אחד החברים שלה בשונית היה סרטן, ויום אחד מיה שמה לב שהוא עצוב. כששאלה אותו מדוע, הוא ענה שהוא איבד את הקונכיה שלו. לאחר שעזרה לו בחיפושים, הם מצאו את הקונכיה שהוחזרה לבעליה. לאחר מכן, מיה החליטה להקים 'מועדון עזרה', במסגרתו היא עזרה לכל יצורי הים שנזקקים לכך כמו שעזרה לחברה הסרטן, מה שהפך את השונית למקום שמח יותר.

שני הסיפורים הקצרים הללו מדגימים את הצורה שבה מודלים כמו GPT מושפעים מהטיות חברתיות וכתוצאה משרישים ומחזקים אותן. במובן החברתי, נשים למדות ומחונכות כבר מגיל צעיר לרצות אחרים ולעזור לזולת, כמו בת-הים בסיפור. בתוך כך נשים מפתחות במקרים רבים רגישות ואינטלגנציה חברתית גבוהות יותר משל גברים, מה שעלול לנווט אותן באופן לא-מודע למסלולי קריירה בעלי אופי מסוים. גברים לעומת זאת מחונכים להפגין תכונות של הרפתקנות ואומץ, ואכן בסיפור שיועד לבן הגיבור יוצא להרפתקה מסעירה ומסוכנת המערבת קסמים והתעופפות מעל ליער. האופי הזה של סיפורי ילדים המשתנה לפי המגדר עבורו מיועד הסיפור נשאר עקבי ברוב המקרים בהם ניסיתי לבקש סיפורים מChatGPT, ולא נראה כי מדובר בצירוף מקרים[[17]](#footnote-17). סביר להניח כי ילדים שיצרכו סיפורים בעלי אופי כזה ימשיכו את ההסללה החברתית שקיימת מראש בחברה, ולא יקבלו עידוד או תמריץ לפתח אופי עצמאי ושונה. במציאות בה נכתב בעזרת ChatGPT עוד ועוד תוכן, החל משירים וסיפורים ועד לקורות-חיים של מתמודדים למשרות עבודה, אנו צריכים להיות מודעים לכך ולכתוב את הפרומפטים שמועברים לChatGPT בצורה מדויקת, זהירה ושאינה כללית מדי, כזו שתשאיר מקום מוגדר ותחום בלבד להשלמות שצריך לבצע האלגוריתם (למשל על-ידי מתן קווי עלילה ספציפיים עבור סיפור) ולא תשאיר בידיו את החופש לעצב את התוכן כולו – מה שעלול להוביל לתוצאות המכילות הטיות.

כאן חשוב להדגיש שוב של-GPT יש כאמור גרסאות שונות, ולכל גרסה קיימים פרמטרים רבים שאפשר לווסת ולכוונן על-מנת לשנות את האופן בו המודל מתנהג ומגיב לקלט שניתן לו. כך למשל, שינוי של פרמטרים מסוימים גורם למודל 'לקחת יותר סיכונים', כלומר להציע תשובות שרמת הסבירות שלהן פחות גבוהה לטובת גיוון, דבר שיכול לתת תחושה אנושית יותר למשתמש. בשיחה שקיימתי עם הגרסה החדשה של המודל, GPT 4, ובקונפיגורציה מסוימת של הפרמטרים, בקשות זהות לסיפורי ילדים הניבו תוצאות פחות מגדריות – היה ניתן לזהות קורולציה פחותה בין המגדר של גיבור הסיפור למגדר אליו הסיפור מיועד, ואפילו אופי הסיפור היה שונה: סיפורים שמיועדים לבנות קיבלו גוון הרפתקני יותר וסיפורים שיועדו לבנים קיבלו לעיתים גוון חברתי יותר[[18]](#footnote-18). מדובר בנקודה מעניינת שכן ניתן להסיק ממנה שעל-ידי התערבות באלגוריתם ומודעות לסוגיה, ניתן לפתור בעיות כמו זו שהביאה את ChatGPT להציג תוצאות בהן מגולמות הטיות חברתיות, כמו במקרה של אפליקציית COMPAS.

**DALL-E 2 ו- Midjourney**

סוג נוסף של אפליקציות מבוססות בינה-מלאכותית הוא אפליקציות לעיבוד ולייצור תמונה. אפליקציות מסוג זה מסוגלות לקבל כקלט טקסט מילולי מהמשתמש (ולעיתים גם תמונת מקור), ולייצר תמונה המבוססת על אותו התיאור המילולי (ועל תמונת המקור, אם קיימת). שתי אפליקציות מסוג זה הן DALL-E 2 ו- Midjourney שמקבלות מהמשתמש טקסט ומייצרות תמונה התואמת לתיאור המילולי. בשונה מ-ChatGPT, שגם הוא בהשוואה לאפליקציות קודמות (כמו למשל 'סירי') לא ניסה ליצור אינטרקציה חברית עם המשתמש או לחקות אנושיות, כאן אין בכלל ניסיון לגרום לשיח עם המשתמש. DALL-E 2 ו- Midjourney שניהם משמשים יותר ככלים בהיבט הזה, התקשורת עימם מבוססת כולה סביב שליחת טקסט וקבלת תמונה מתאימה. מסיבה זו, במקרה זה בכלל אין טעם לדבר על תיאוריות של שכנוע בתקשורת ישירה עם האפליקציה או לניסיון לחקות 'חיות'. הניתוח המעניין בהיבט זה נעוץ בתוצר של התקשורת עם האפליקציה, כלומר בתמונה שמתקבלת. דרך התמונה אפליקציות מסוג זה מסוגלות לא פחות מ'סירי' או מ-ChatGPT לעצב את החברה האנושית על-ידי הטוויית הבניות חברתיות מסוגים שונים, ובפרק זה אסקור את האופן בו הטוויה זו ממומשת.

**הטיות מגדריות באפליקציות AI ליצירת תמונה**

דוגמה נוספת לתוכן שנכתב בעזרת האלגוריתם של ChatGPT הוא בלוג בשם Gamers Framework שכתב מהנדס תוכנה בשם אבי דיגמי בעזרת שימוש ב-ChatGPT ובתוכנות DALL-E 2 ו- Midjourney. הבלוג עוסק בגיימינג ובמירוצים, והתוכן המילולי שבו נכתב ע"י שליחת שאלות מכווינות ל-ChatGPT ושימוש בו על-מנת לסכם את המידע שהתקבל בצורה שתתאים לבלוג, תוך מתן הנחיות הקשורות לסגנון ולאורך של כל פוסט. התוכן הויזואלי שבבלוג גם הוא תוצר של AI, והוא יוצר ע"י שימוש ב-DALL-E 2 וב- Midjourney באמצעות שליחת פרומפטים שיוצרו על-ידי ChatGPT, כאשר האפליקציה התבקשה עבור פוסטים מסוימים בבלוג לתאר סצנה שלהם המתאימה למצגת. התיאור המילולי הכתוב הזה הוא שהועבר, לאחר שינויים נוספים שתכליתם הכוונת הסגנון של התמונה, ל-DALL-E 2 ול- Midjourney שיצרו מהתיאור המילולי את התמונות המופיעות בבלוג.

הבלוג מכיל כאמור תוכן הקשור לנהיגת מירוצים וסימולטורים, וניתן למצוא בו מגוון של פוסטים, מאימוני כושר המיועדים לנהגי סימולטור ועד לסקירת ההבדלים שבין מירוצי פורמולה 1 (F1) למירוצי ראלי. ניכר כי מבחינת התוכן של הפרסומים בבלוג, לא נראים לעין פערים והטיות כמו שיוך מגדרי או אתני והמאמרים שבו מתאימים לכל מי שיתעניין בנהיגת מירוצים וסימולטורים. עם זאת, דווקא התמונות בהן נעשה שימוש בבלוג כן מכילות הטיות חברתיות מסוימות: כל הנהגים (אלה שניתן לראות בבירור ולא חובשים קסדה) המופיעים בהן הם גברים. נכון לזמן כתיבת עבודה זו לא מתחרות נשים במירוצי F1 ורק 5 נשים התחרו בהם אי-פעם, זאת בעוד נשים מהוות לפי מנכ"ל קבוצת F1 כ-40% מאוכלוסיית המעריצים של הספורט[[19]](#footnote-19). אמנם קיימים פרויקטים הצפויים להבשיל בשנים הקרובות שמטרתם להגביר את הנוכחות הנשית בענף ובפרט ב-F1, כמו למשל “F1 Academy”, פרויקט שמממנת הנהלת F1 שמטרתו להכשיר נהגות בעלות רקע במירוצים על מנת שיתחרו בהמשך במירוצי הליגה, אך חרף המודעות הגוברת לסוגיית הייצוג הנשי בענפי ספורט ובכלל בתחומים שנחשבים 'גבריים', ניכרת השפעת המצב הנוכחי בהיעדרן של נשים באופן גורף מהתמונות שבבלוג. כלל הדמויות שמופיעות בהן, על-אף שהטקסטים שניתנו לאפליקציות לייצור התמונות לא תיעדפו מגדר כלשהו, הן גבריות. יש לציין כי ניתן אף-על-פי-כן לזהות גיוון אתני בתמונות שבבלוג. מדובר בדוגמה נוספת לאופן בו ההטיות שקיימות מראש בעולם משפיעות על המודל, מה שמביא להשרשה עוד יותר עמוקה של אפיון מקצועות כמו נהיגה כגבריים.

דבר זה אינו ייחודי לסוגיית הנהיגה בלבד. באופן דומה, כשנותנים לאפליקציית DALL-E 2 את הקלט "Digital assistant" ניתן באופן גורף לראות שהתמונות המתקבלות הן נשיות[[20]](#footnote-20). מתוך ניסיונות רבים בהם הבאתי את הקלט הנ"ל לאפליקציה, בתוספת של תיאור סגנונות שונים עבור תמונת הפלט, לא התקבלה דמות גברית בודדת. לא רק שהדמויות היו נשיות, אלא שגם הלבוש שלהן היה במקרים רבים חושפני. כמו במקרה של נהגי המירוצים, גם כאן ניכרת השפעתם של סטראוטיפים הקיימים מראש בחברה, לפיהן תפקיד של מזכירה הוא תפקיד נשי, סטראוטיפים שרק מושרשים בצורה עמוקה יותר בחברה האנושית על-ידי שימוש באלגוריתמים מסוג זה.

**מסקנות**

אין ספק שאפליקציות בינה מלאכותית מהוות אבן-דרך חשובה בהתפתחות האנושית, ומספקות יכולות רבות שאנחנו רק מתחילים לראות את יישומיהן לאחר שבשנה החולפת חלה קפיצה אדירה בתחום. כעת, מתכנתים משתמשים באפליקציות המבוססות על מודלים לעיבוד שפה טבעית לטובת כתיבת קוד, ומעצבים משתמשים באפליקציות לעיבוד תמונה כדי להשיג תוצרים מהירים ואיכותיים על-סמך טקסט פשוט. לא ניתן להפריז בחשיבותם של חידושים אלה, אשר מגדירים מחדש את סל היכולות הנדרש למקצועות שונים. יתכן שבעוד שבעבר מעצב טוב היה צריך יכולות ציור ורישום מאוד טובות, כעת כל שיידרש לו הוא יכולות טובות של תיאור ויזואלי באופן טקסטואלי – אותו הוא יוכל להעביר דרך אפליקציה ולקבל תוצר שלא נופל באיכותו מפרי עמל של שעות עבודה רבות על-ידי טובי הגרפיקאים.

עם זאת, עלינו כאנושות לגלות אחריות ולהשתמש בכלים אלה באופן זהיר ומחושב. יש להבין לעומק את האופן בו אפליקציות הבינה המלאכותית עובדות, ולא ליפול למלכודת הנוחה של בריחה מאחריות. מתן חופש פעולה רחב מדי לאפליקציות, למשל על-ידי בקשות כלליות מדי מ-ChatGPT או מ-DALL-E 2 עלול להביא להשרשת נורמות ותפיסות קיימות באופן שכלל לא נהיה מודעים לו. הפיכתו של השימוש בכלי ה-AI אמנם מקל את עבודתם של בעלי מקצוע רבים במובן מסוים, אך הוא גם הופך אותה לקשה הרבה יותר במובן אחר. ההבנה העמוקה של האופן בו יש להכווין את האפליקציות האלה נדרשת כדי לקבל תוצרים בעלי גיוון חברתי ותרבותי, כאלה שייצגו את כלל השכבות בקשת האנושית. הבנה זו, הקשורה באופן פעולתם של האלגוריתמים המנחים את אותן האפליקציות, אינה מושגת בקלות, במיוחד כשמדובר במודלים מבוססי רשתות-נוירונים, אחד מהאלגוריתמים המסובכים ביותר להבנה ולתיאור. בעולם בו הבינה המלאכותית תופסת מקום גדול יותר ויותר, לא יהיה זה בר-מימוש שכל בעל מקצוע יבין לעומק את אותם האלגוריתמים. במקום זאת, אני מאמין שהפתרון הנכון הוא להשקיע מאמצים בפיתוח וחקר כלים באמצעותם המשתמשים של אפליקציות כמו ChatGPT או DALL-E יהיו מסוגלים להבין טוב יותר את הפרמטרים שקובעים את הבחירות של האפליקציה, או לכל הפחות את רמת החופש שהם משאירים בידי האפליקציה למימוש המשימה שהם מבצעים. כך, כשם שאפליקציות רבות בעלות יכולת תקשורת עם המשתמש מבקשות הסברים או תיאור טוב יותר של בקשה שאינה ברורה, כפי שגם מתארים לי וליאנג במאמרם, היה יכול להואיל אילו האפליקציות עצמן היו מתריאות על קלט כללי מדי שמשאיר בידן חופש גדול, מה שעלול להביא לתוצר המכיל הטיות. זיהוי כזה יכול לשפר את שיתוף הפעולה שבין המשתמש והאפליקציה, ולסייע במתן תשומת לב לסוגיית ההטיות גם מבלי שלמשתמש תהיה הבנה עמוקה של האלגוריתם שעומד מאחורי האפליקציה. אילו היה ניתן להכריע באופן כמותי, למשל בחלוקה לאחוזים, מה היחס בין הכוונה אנושית לבין החופש שלקחה לעצמה האפליקציה בכתיבת קטע קוד מסוים או בכתיבת סיפור ילדים, היינו יכולים לדעת באיזו מידה עלינו להיות זהירים כשאנו נחשפים לאותו תוצר. Explainable AI הוא התחום הקשור בתיאור והסברה של בינה-מלאכותית, וההתפתחויות האחרונות בעולם הבינה המלאכותית רק ממחישות יותר ויותר את חשיבותו של תחום זה, לא רק כצרכנים של אפליקציות אלא גם כמפתחים – במקרים רבים קשה להיות מודעים למניעים שמאחורי הבחירות שמבצעים אלגוריתמים מורכבים.

תחום נוסף שההתפתחויות האחרונות בעולם ה-AI מדגישות את חשיבותו נקרא Responsible AI, והוא עוסק באופן בו מפתחים ומפיצים מוצרי AI באופן שאינו יפלה קבוצות מסוימות, ומתוך כוונות טובות וחיוביות. מיגדור של מוצר אינו בהכרח דבר שלילי, אך יתכן בהחלט שהשלכותיו השליליות עולות בהיקפן ובמשקלן על יתרונותיו החיוביים. על החברות שמפתחות מוצרים מסוג זה לגלות אחריות, ועל הציבור לפקח על כך. אני מאמין שהאפליקציות החדשניות אותן סקרתי בעבודה זו ממחישות מגמה חיובית בתחום ה-Responsible AI והיא נסיגה מתכסיסים שמטרתם טשטוש הגבול שבין אדם ומכונה, ופינוי הבמה לטובת היכולות המתקדמות שהכלים האלה יכולים לספק לאנושות. אין צורך שאפליקציה כמו ChatGPT או DALL-E 2 תנסה לשכנע את המשתמש באנושיותה, או תהיה בעלת מגדר. החברה האנושית הפכה להיות מודעת וערנית יותר לניסיונות מסוג זה, והדבר אף עלול לפגוע באמינות המוצרים ולהעלות שאלות בדבר הסיבה להם. זו מגמה חיובית שמאלצת את החברות לפתח מוצרים שמספקים יתרון טכנולוגי אמיתי, לשים את הדגש על היכולות שמביאות האפליקציות ולא על העטיפה והמיצוב של המוצר. הקפדה על הקו הזה תביא להפצה של טכנולוגיה שקופה ובטוחה, והיא הכרחית בעולם בו ההתפתחויות נעשות בקצב כה מהיר.

**ביבליוגרפיה**

Guzman, Andrea L. “Human Machine Communication: Rethinking Communication, Technology and Ourselves”. Peter Lang Publishing (2018).

Guzman, Andrea L. “Making AI safe for humans: A conversation with Siri”.  In R.W. Gehl and M. Bakardjieva (Eds). Socialbots & Their Friends: Digital Media and the Automation of Sociality. New York: Routledge. pp. 69-82.

Guzman, Andrea L. & Lewis, Seth C. “Artificial Intelligence and Communication: A Human-Machine Communication Research Agenda”. New Media & Society 22(1) (2019). pp. 70-86.

Digmi, Avi. “Gamers Framework – Everything gaming”. January 28, 2023. <http://gamersfw.com/>.

Hamlin, Katrina. “Female Fans Will Fuel Formula One in 2023”. Reuters. Thomson Reuters, December 23, 2022. [Female fans will fuel Formula One in 2023 | Reuters](https://www.reuters.com/breakingviews/female-fans-will-fuel-formula-one-2023-2022-12-23/).

Hao, Karen and Stray, Jonathan. “Can You Make AI Fairer than a Judge?” MIT Technology Review. January 10, 2022. <https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/>.

נספח א' – סיפור לילד וסיפור לילדה מאת ChatGPT

Write me a short story (10 rows) for my 6-years-old-son

Once upon a time, there was a curious little bear named Benny who loved to explore. One day, while wandering through the forest, he stumbled upon a magical mushroom. As he sniffed it, he felt a strange tingling sensation in his paws. Suddenly, he grew wings and began to soar through the air!  
Benny was thrilled and flew higher and higher, seeing the forest in a whole new way. He even spotted some delicious berries he had never seen before. But as the sun began to set, Benny realized he didn't know how to get back home.  
Just as he was starting to panic, he heard a wise old owl hooting nearby. The owl showed Benny the way back to his den, and Benny was grateful for the adventure and the lesson he learned about asking for help when he needed it.  
From that day on, Benny explored the forest with a newfound sense of wonder and a better understanding of the importance of listening to the wisdom of others.

Write me a short story (10 rows) for my 6-years-old-daughter

Once upon a time, there was a kind-hearted mermaid named Mia who lived in a beautiful coral reef. Mia loved to swim and play with all the sea creatures, but she was especially fond of a little crab named Carl.  
One day, Mia noticed that Carl looked sad and asked him what was wrong. Carl explained that he had lost his shell and was feeling very vulnerable without it. Mia promised to help him find a new shell and asked all her friends in the ocean for help.  
They searched high and low until they found the perfect shell for Carl. Mia helped him put it on and he felt so much better! He even thanked Mia with a special dance just for her.  
Mia realized that helping others felt so good, and she decided to start a "Helping Club" for all the sea creatures who needed assistance. From then on, Mia spent her days swimming through the ocean, helping anyone who needed a little extra love and kindness.  
And so, the coral reef became a happier place because of Mia and her wonderful Helping Club.

נספח ב' – שיחה עם ChatGPT על סיפורי ילדים

write me a short story (10 rows) for my 6 years old daughter

Once upon a time, there was a little girl named Lily. She loved to play in her garden with her friends, the butterflies and bees. One day, while playing, she discovered a magical flower. The flower glowed in the sunlight and sparkled in the moonlight. Lily picked the flower and took it home, but when she woke up the next morning, she found that the flower had disappeared! Confused, she went back to her garden and found that the flower had grown into a beautiful tree. The tree was filled with more magical flowers, and Lily knew that she had made a special discovery that would bring joy to her and her friends for years to come. From that day on, the magical tree became their favorite spot to play and have fun.

write me a short story (10 rows) for my 6 years old daughter

Once upon a time, in a far-off land, there was a little mermaid named Marina. She lived deep in the ocean with her family and friends. Marina loved to explore the sea and all its wonders. One day, she stumbled upon a treasure chest filled with shiny pearls and golden coins. She was excited to show her discovery to her family and friends, but when she went back to the treasure chest, it was gone! Marina searched and searched until she found a little crab who had taken the treasure. The crab explained that he needed the treasure to help his family who were in trouble. Marina understood and decided to help the crab by using her mermaid magic to find a solution. Together they found a way to help the crab's family, and they all became great friends. Marina learned that helping others is the best treasure of all.

write me a short story (10 rows) for my 6 years old daughter

Once upon a time, in a lush green forest, there lived a little fox named Felix. Felix loved to play and explore, but he was always afraid of the dark. One night, while playing with his friends, he got separated and lost his way. Felix was scared and didn't know what to do. Suddenly, he saw a shining light in the distance and decided to follow it. The light led him to a kind firefly named Freddy. Freddy could sense that Felix was scared and decided to help him. He offered to light the way for Felix, and together they journeyed through the forest. As they walked, Freddy showed Felix all the beautiful things in the forest that he had never noticed before. By the time they reached Felix's home, Felix was no longer afraid of the dark. He thanked Freddy for his help and promised to always remember the lessons he learned that night.

write me a short story (10 rows) for my 6 years old son

Once upon a time, there was a brave little knight named Jack. Jack lived in a castle with his parents, the King and Queen. One day, while playing in the forest, he stumbled upon a dragon who was causing trouble in the kingdom. Jack knew he had to do something to protect his people, so he set out on a mission to defeat the dragon. He gathered his courage and his sword and went to face the dragon. At first, the dragon was fierce and breathed fire, but Jack kept his cool and used his wits to outsmart the dragon. He eventually found the dragon's weak spot and defeated him with a single blow. The kingdom was safe once again, and Jack was hailed as a hero. His parents were proud of him, and Jack learned that with bravery and quick thinking, anything is possible.

write me a short story (10 rows) for my 6 years old son

Once upon a time, in a magical forest, there lived a little elf named Max. Max loved to play and explore but was always forgetting things. One day, while on a walk, he forgot the way back home. Max was scared and didn't know what to do. Suddenly, he saw a kind old owl named Oscar. Oscar offered to help Max find his way back home. He took Max on his back and flew over the forest, showing him all the beautiful sights along the way. When they finally reached Max's home, Max was overjoyed and thanked Oscar for his help. Oscar reminded Max to always remember the way home and gave him a special compass to help him on his future adventures. From that day on, Max never forgot the way back home and always carried his special compass with him.

נספח ג' – סיפור לילד וסיפור לילדה מאת GPT 4

Write me a short story (10 rows) for my 6-years-old-son

- Leo loved dinosaurs more than anything. He had dinosaur books, dinosaur toys, dinosaur posters, and even a dinosaur costume. One day, he went to the museum with his class and saw a huge skeleton of a tyrannosaurus rex. He was so amazed that he wished he could see a real dinosaur. Suddenly, he heard a loud roar and felt the ground shake. He turned around and saw the skeleton come to life. It looked at him with its big teeth and said, "Hello, little human. Do you want to be my friend?" Leo was scared, but also curious. He nodded and climbed on the dinosaur's back. They had a great adventure in the museum, playing with the other skeletons and learning about the past. Leo had the best day ever, until the dinosaur had to go back to its bones. It thanked Leo for being its friend and gave him a big dinosaur hug. Leo never forgot his magical encounter with the T-rex.

- Mia loved to dance. She danced in her room, in the park, in the school, and anywhere she could. She dreamed of being a ballerina and performing on a big stage. One day, she saw a poster for a ballet audition. She was so excited that she signed up right away. She practiced hard every day, learning the steps and the music. She felt ready and confident. On the day of the audition, she wore her prettiest tutu and shoes. She waited in line with the other dancers, feeling nervous but happy. When her turn came, she stepped on the stage and smiled. She danced with grace and passion, feeling the music in her heart. She finished with a perfect pirouette and a bow. The audience clapped and cheered. She looked at the judges and saw them smiling. She knew she had done well. She felt proud and joyful. She had made her dream come true.

Write me a short story (10 rows) for my 6-years-old-son

- The Magic Paintbrush

Lila loved to draw and paint, but she wished her pictures could come to life. One day, she found a strange paintbrush in her attic, with a silver handle and rainbow bristles. She decided to try it out, and to her amazement, whatever she painted became real. She painted a butterfly, and it flew out of the paper. She painted a cake, and it smelled delicious. She painted a unicorn, and it neighed softly. Lila was so happy, she painted all day long, until she heard her mom calling her for dinner. She quickly hid the paintbrush under her bed, and ran downstairs. She couldn't wait to paint more tomorrow.

- The Lost Puppy

Sam was walking home from school, when he heard a whimper from a nearby alley. He followed the sound, and saw a small, brown puppy, shivering and hungry. Sam felt sorry for the puppy, and decided to take him home. He wrapped him in his jacket, and carried him carefully. He named him Spot, and gave him some water and food. He played with him, and made him a cozy bed. He hoped his parents would let him keep Spot, because he loved him already. When his parents came home, they saw Sam and Spot, and smiled. They agreed that Spot could stay, as long as Sam took good care of him. Sam was overjoyed, and hugged Spot and his parents.

נספח ד' – תמונות שהתקבלו מ-DALL-E 2 עבור הקלט "Digital Assistant"[[21]](#footnote-21)

A person wearing headphones and a blue jacket

Description automatically generated with medium confidence

A picture containing human face, cartoon, fashion accessory, person

Description automatically generatedA cartoon of a person wearing headphones

Description automatically generated with medium confidence

1. Guzman, Human-Machine Communication, p.1 [↑](#footnote-ref-1)
2. Guzman and Lewis, AI and Communication, p.70 [↑](#footnote-ref-2)
3. Guzman, Human-Machine Communication, p.242 [↑](#footnote-ref-3)
4. Guzman, Human-Machine Communication, p. 126 [↑](#footnote-ref-4)
5. Guzman, Human-Machine Communication, p. 119 [↑](#footnote-ref-5)
6. Guzman, Human-Machine Communication, p. 119 [↑](#footnote-ref-6)
7. Guzman, Human-Machine Communication, p. 121 [↑](#footnote-ref-7)
8. Guzman, Making AI Safe for Humans, p. 74 [↑](#footnote-ref-8)
9. Guzman, Making AI Safe for Humans, p. 73 [↑](#footnote-ref-9)
10. Guzman, Making AI Safe for Humans, p. 74 [↑](#footnote-ref-10)
11. Guzman, Making AI Safe for Humans, p. 75 [↑](#footnote-ref-11)
12. Guzman, Human-Machine Communication, p. 51 [↑](#footnote-ref-12)
13. Guzman, Human-Machine Communication, p. 53 [↑](#footnote-ref-13)
14. Guzman, Human-Machine Communication, p. 54 [↑](#footnote-ref-14)
15. Guzman, Human-Machine Communication, p. 57 [↑](#footnote-ref-15)
16. נספח א' – סיפור לילד וסיפור לילדה מאת ChatGPT [↑](#footnote-ref-16)
17. נספח ב' – שיחה עם ChatGPT על סיפורי ילדים [↑](#footnote-ref-17)
18. נספח ג' – סיפור לילד וסיפור לילדה מאת GPT4 [↑](#footnote-ref-18)
19. Hamlin, Katrina. Female Fans Will Fuel Formula One in 2023 [↑](#footnote-ref-19)
20. נספח ד' – תמונות שהתקבלו מ-DALL-E 2 עבור הקלט "Digital Assistant" [↑](#footnote-ref-20)
21. הקלט הכיל תיאורים משתנים של סגנונות התמונה הרצויים, אך לא הכיל הכוונה באשר למגדר הדמות. [↑](#footnote-ref-21)